Northeastern

Department of Civil and
Environmental Engineering
400 Snell Engineering
360 Huntington Ave.
Boston, MA 02115

CIVE 7381 Transportation Demand
Problem Set #1
Due: Wednesday, September 18, 2024

A planning organization has collected data on **all** families (entire population!) vacationing in a particular area during the summer. The entire population consists of 200 families with the associated information contained in the file PS1-data.xls, which is available on Canvas Problem Set 1. The file contains the following information for each family.

Duration of vacation	*DUR*	1 if vacation is longer than a week 0 otherwise
City	*CITY*	1 if vacation is at a city 0 otherwise
Camping	*CAMP*	1 if camping during vacation 0 otherwise
Income	*INC*	Average disposable income of household to which traveler belongs (in dollars)
Expenditure	*EXPEND*	Average amount spent by the family during the vacation (in dollars)
Education	*EDUC*	Education, measured by number of years after high school of the highest degree in household

1. Using the entire population perform the following tasks
    a. Calculate the means and standard deviations for *EXP, DUR, INC,* and *EDUC*.
    b. Calculate the correlation and covariance matrix for these four variables.
    c. Plot histograms for these variables.
    d. Plot the scatter plot matrix for these variables.
2. Using the *sampling* function in Excel generate 5 random samples from the population, each of size 40.
    a. For each sample find the sample mean of the various attributes and show them in a table.
    b. Plot the frequency histogram of the sample mean values for Income.
    c. Find the average and variance of the sample mean for income (across the five samples). Compare the average to the true population mean and the variance to the variance calculated by the formula for the standard error.
3. Generate two new samples: Sample 1 is of size 100 (sample 100 observations) and Sample 2 consists of 175 observations. Repeat parts 1a, 1b and 1c. Comment on how the sample averages and histograms compare to the corresponding quantities in the entire population. Which sample is closer? Comment on the impact of the sample size.

4. Using the data from one sample generated in (2) and Sample 2 generated in (3) develop the confidence interval for the mean of each sample at the 95% level. Compare the corresponding intervals and comment on the effect of the sample size.
5. Using the data from Sample 2 (question 3), test the null hypothesis H_0: μ= \$25,000, against the alternative hypothesis H_1: $\mu \neq$ \$25,000, where μ is the population mean (make any assumption regarding the variance that you think appropriate).
6. Based on the results of the analysis in questions 2-4 what is the sample size you would recommend (obviously larger samples are more accurate but also more expensive)?
7. Based on your understanding of the data and the problem, which of the variables above you would use for developing a model to predict expenditure per family?
8. What other variables, not included in the data set, do you consider important in building a better model to predict expenditures per family?

You can use Excel to generate the various random samples. First make sure that the Analysis Tool add-inn is activated (add it by File → Options → Add-ins). This will allow us to use a Random Number Generator. There are different ways to draw a random sample from a data set. See for example:

WikiHow. How to Create a Random Sample in Excel, http://www.wikihow.com/Create-a-Random-Sample-in-Excel

SurveyMonkey Blog. How to Create A Random Sample in Excel, https://www.surveymonkey.com/mp/random-sample-in-excel/